Received: 20 March 2019 | Revised: 10 May 2019 | Accepted: 13 May 2019

DOI: 10.1111/1755-0998.13044

RESOURCE ARTICLE

MOLECULAR ECOLOGY RESOURCES WILEY

Rapid identification and interpretation of gene–environment associations using the new R.SamBada landscape genomics pipeline

Solange Duruz[1] | Natalia Sevane[2] | Oliver Selmoni[1] | Elia Vajana[1] | Kevin Leempoel[3] | Sylvie Stucki[1] | Pablo Orozco-terWengel[4] | Estelle Rochat[1] | Susana Dunner[2] | The NEXTGEN Consortium | The CLIMGEN Consortium | Michael W. Bruford[4] | Stéphane Joost[1]

[1]Laboratory of Geographic Information Systems (LASIG), School of Architecture, Civil and Environmental Engineering (ENAC), Ecole Polytechnique Fédérale de Lausanne (EPFL), Lausanne, Switzerland

[2]Departamento de Producción Animal, Facultad de Veterinaria, Universidad Complutense de Madrid, Madrid, Spain

[3]Department of Biology, Stanford University, Stanford, California

[4]School of Biosciences, Cardiff University, Cardiff, Wales, UK

Correspondence
Stéphane Joost, Laboratory of Geographic Information Systems (LASIG), School of Architecture, Civil and Environmental Engineering (ENAC), Ecole Polytechnique Fédérale de Lausanne (EPFL), Lausanne, Switzerland.
Email: stephane.joost@epfl.ch

Funding information
FACCE ERA-NET Plus, Grant/Award Number: ANR-14-JFAC-0002-01; H2020 Marie Skłodowska-Curie Actions, Grant/Award Number: DLV-655100; Biotechnology and Biological Sciences Research Council, Grant/Award Number: BB/M019276/1; FP7 Food, Agriculture and Fisheries, Biotechnology, Grant/Award Number: 244356

Abstract

SAMβADA is a genome–environment association software, designed to search for signatures of local adaptation. However, pre- and postprocessing of data can be labour-intensive, preventing wider uptake of the method. We have now developed R.SamBada, an R-package providing a pipeline for landscape genomic analysis based on SAMβADA, spanning from the retrieval of environmental conditions at sampling locations to gene annotation using the Ensembl genome browser. As a result, R.SamBada standardizes the landscape genomics pipeline and eases the search for candidate genes of local adaptation, enhancing reproducibility of landscape genomic studies. The efficiency and power of the pipeline is illustrated using two examples: sheep populations from Morocco with no evident population structure and Lidia cattle from Spain displaying population substructuring. In both cases, R.SamBada enabled rapid identification and interpretation of candidate genes, which are further discussed in the light of local adaptation. The package is available in the R CRAN package repository and on GitHub (github.com/SolangeD/R.SamBada).

KEYWORDS
gene-environment association, landscape genomics, Lidia cattle breed, local adaptation, Moroccan sheep, R-package

The NEXTGEN Consortium: https://nextgen.epfl.ch

The CLIMGEN Consortium: https://climgen.bios.cf.ac.uk/

1 | INTRODUCTION

Local adaptation implies the existence of advantageous alleles conferring a population living in its native habitat a higher fitness than any other allochthonous population living in the same habitat (Kawecki & Ebert, 2004). Landscape genomics methods (Joost et al., 2007), including genome–environment association (GEA), are among the approaches used to detect signatures of local adaptation and have become increasingly popular, mainly due to the decreasing cost of sequencing, but also because of the recent availability of fine-scale environmental data sets (Balkenhol et al., 2019; Rellstab, Gugerli, Eckert, Hancock, & Holderegger, 2015). However, the massive amount of data that can be analysed due to these improvements have made the development of more efficient tools essential (Stucki et al., 2017).

To this end, SAMβADA was developed to perform large amounts of logistic regressions between genetic markers and multiple environmental variables (Stucki et al., 2017). SAMβADA computes uni- or multivariate models between a binary genetic variable (e.g., the presence/absence of a genotype) and one or more environmental variables. Significance is assessed against a null model (i.e., constant model in the case of univariate or a parent model in the multivariate case). Population structure can be accounted for by treating one or several population variables as environmental variables in multivariate analysis. SAMβADA is written in C++ with a particular emphasis on high-performance computing (HPC). Since its publication, SAMβADA, as applied alone or in combination with other methods, proved useful to target putative genomic regions underlying local adaptation in a wide variety of species, including domestic animals such as swine and cattle (Cesconeto et al., 2017; Vajana et al., 2018), wild animals such as the freshwater sculpin and European pond turtle (Lucek, Keller, Nolte, & Seehausen, 2018; Pereira, Teixeira, & Velo-Antón, 2018), and many different plant species including the European beech and the cow-tail fir (Cuervo-Alarcon et al., 2018; Shih, Chang, Chung, Chiang, & Hwang, 2018).

Despite its many advantages, SAMβADA's command-line format is sometimes laborious and the amount of pre- and postprocessing represents an obstacle to its widespread use. Indeed, a typical processing chain, such as the one proposed by Stucki et al. (2017), includes (a) the use of a GIS software to retrieve environmental information at sampling locations; (b) molecular data filtering by standard software such as PLINK (Chang et al., 2015); and (c) the inclusion, whenever present, of population structure usually computed with a dedicated software such as ADMIXTURE (Alexander, Novembre, & Lange, 2009). Similarly, postprocessing of results involves (a) the computation of *p*- or *q*-values (Storey, 2003) for the association tests involving each genotype; (b) the production of maps and plots (typically Manhattan plots) in which the location in the genome (i.e., the position in base pair) of a point representing the result of a model is difficult to establish since the plot is rarely interactive; (c) the formulation of queries to the Ensembl genome browser (Hubbard et al., 2002) to search for candidate genes adjoining the single-nucleotide polymorphisms (SNPs) highlighted.

However, the R software (R Core Team, 2018) provides an open-source computing environment adapted to different fields in Biology, in which many of the above-mentioned pre- and post-processing tasks can be found in various R-packages. Further, R can be coupled with compiled languages (such as C++) so as to be more efficient when processing large data sets (see e.g., the case of the software LFMM 2; Caye, Jumentier, Lepeule, & François, 2019, p. 2).

In this context, we developed R.SamBada, an R-package designed to facilitate and enhance the whole data process described above by integrating multiple existing packages and building new functions into one easy-to-use pipeline. We present the use of the package by illustrating its benefits with two case studies for which driven signatures of selection were investigated as part of the ClimGen project (https://climgen.bios.cf.ac.uk/). The first data set consists of 160 Moroccan sheep genotyped with whole genome sequencing (WGS) and characterized by no clear population structure, while the second one encompasses a Spanish Lidia Cattle population of 349 samples genotyped with 50 K SNP chip, with one population variable. Both data sets are already published (see Data availability section) but have not yet been analysed with SAMβADA.

2 | MATERIALS AND METHODS

We first present R.SamBada, with an overview of its functions, and then describe its application to two case studies from the ClimGen (https://climgen.bios.cf.ac.uk/) project, detailing how the genetic data were collected and prepared for subsequent analyses. Both studies investigate climate-mediated selection at the genome level: the first analysis is carried out on a Moroccan sheep data set using whole genome sequences, and the second one involves a Spanish cattle breed (Lidia) genotyped with the Illumina BovineSNP50 array.

2.1 | Implementation

R.SamBada provides functions for (a) preparing the genetic (i.e., SNPs) and environmental information to be processed (preprocessing), (b) running SAMβADA directly into the R environment (processing) and (c) performing post hoc analyses on the basis of SAMβADA's output (postprocessing). The following sections detail these different steps (Figure 1).

2.1.1 | Preprocessing

Three functions have been implemented to perform the main operations required before running SAMβADA. First, *prepareGeno* is used to prepare the genomic file, by treating a SNP input data set from various formats (.vcf, .gds, .ped or .bed) and generating a filtered file complying with SAMβADA's input standards. *prepareGeno* relies on the SNPRELATE package (Zheng et al., 2012) to perform standard quality control (QC) for minor allele frequency (MAF), linkage disequilibrium (LD) and missingness. In order to assist users in selecting adequate pruning levels, *prepareGeno* displays the frequency distributions of

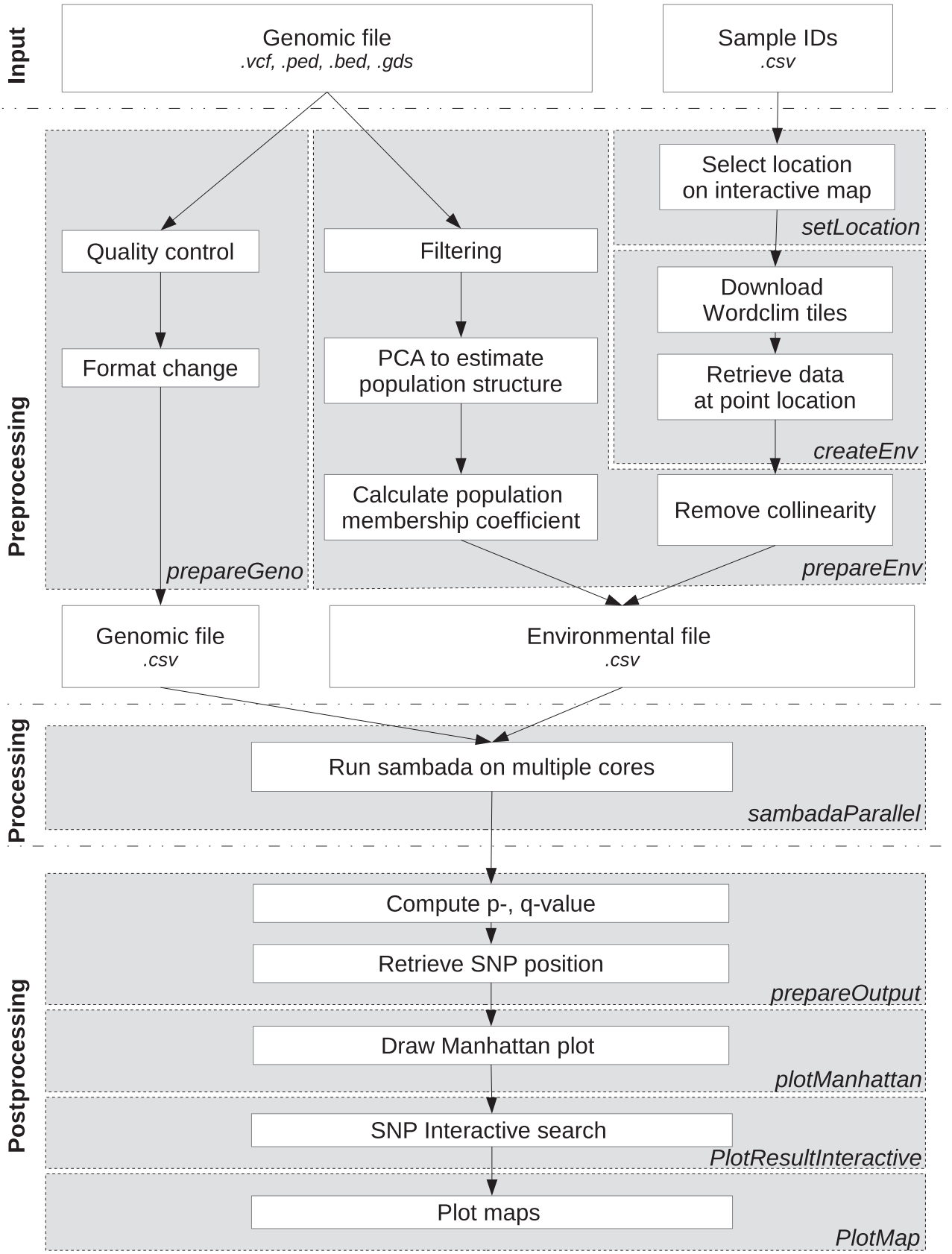

FIGURE 1 Overall functionalities and process in R.SamBada. Grey boxes with italic names indicate functions included in the package. The process starts with a genomic file and a file with sample locations or list of IDs. The preprocessing will format the genomic file and prepare the environmental file; SAMβADA is then run parallelly on multiple cores; after computing of *p*-, *q*-values, Manhattan plots and maps can be drawn and Ensembl database can be queried

MAF, LD and missingness along with the proportion of SNPs discarded corresponding to the thresholds applied; in this way, QC can be tailored to avoid reducing the data set too much while controlling for missing information.

Second, if coordinates are not available, *setLocation* can be used to open a local web page that assist users in defining sample locations using mouse-clicks on an interactive map. The projection system used is WGS84 (corresponding EPSG – European Petroleum Survey Group – code: 4326), a worldwide system with coordinates in degrees (longitude/latitude) (more information on projections in Leempoel et al., 2017).

Then, *createEnv* provides the user with a pipeline to produce an environmental data set out of the file containing sample locations. If raster files representing environmental variables are available, then habitat information is directly derived at the sampling locations. However, if these files are not present, *createEnv* is able to use the samples' geographic coordinates to identify the correct tiles in the WorldClim (Hijmans, Cameron, Parra, Jones, & Jarvis, 2004) and SRTM (Shuttle Radar Topography Mission) (Farr et al., 2007) databases and to download adequate climatic and altitudinal information. The WorldClim database contains monthly minimum, maximum and average temperature and total precipitation together with a series of bioclimatic variables computed from these variables (e.g., precipitation of wettest quarter of the year, complete list available at http://www.worldclim.org/bioclim), while SRTM only provides altitude. Coordinates can be given in any projection system (as long as the EPSG code of the projection is given as an input parameter of the function). A comma-separated value (.csv) file is then returned containing the sample IDs, their locations and the values of the corresponding environmental variables. The interactive mode shows maps of sample locations, so as to locate potentially misplaced points or erroneously-set projection systems. This function can save substantial effort, since one single command substitutes a long processing chain that typically includes the download of voluminous data for the entire globe, the import of both sample locations and raster environmental data into GIS software and the retrieval of environmental values at point location.

Finally, the *prepareEnv* function produces a file containing the design matrix that SAMβADA will process. At first, highly correlated environmental variables are removed according to a correlation coefficient threshold defined by the user in order to keep only independent eco-climatic factors in the analysis. The interactive mode will show the graph of the number of variables discarded as a function of the chosen correlation threshold. Then, the genetic structure of populations is assessed by means of a principal component analysis (PCA) as implemented in SNPRELATE. The user is provided with the possibility of further processing PCA output by a clustering algorithm, which calculates individual membership coefficients as a function of the distance from the clusters centroids (Lee, Abdool, & Huang, 2009). Changes in the clustering solution according to the chosen k-number of clusters can be interactively visualized. After ordering individuals according to their identifiers (as in the genomic file and necessary for SAMβADA's analysis), a final.csv file is generated, containing the samples' IDs, the retained environmental variables and either the PCA score(s) or the membership coefficient(s) representing population structure.

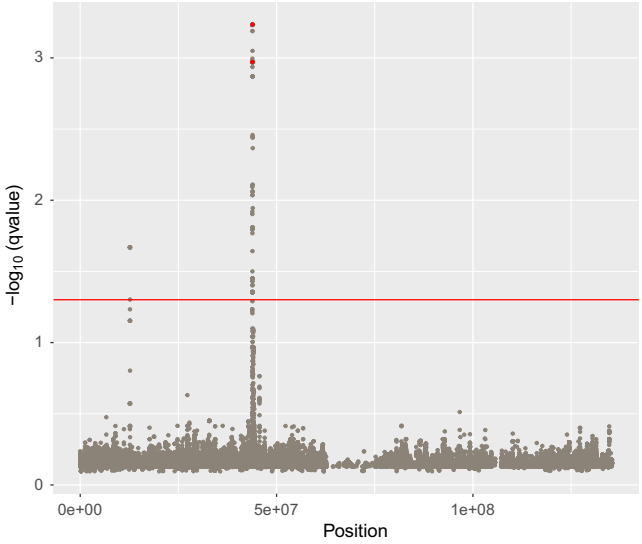

FIGURE 2 Manhattan plot showing the *q*-values for each marker (with G- or Wald-Score > 6) on chromosome 23 of Moroccan sheep associated with annual precipitation as calculated in SAMβADA in a univariate mode. Points in red correspond to models involving two nonsynonymous SNPs (ss1208941124 and ss1208941157) in the MC5R gene (ss1208941124 having the lowest q-value of the two). The red horizontal bar shows a significance threshold of 0.05 [Colour figure can be viewed at wileyonlinelibrary.com]

2.1.2 | Processing

SAMβADA includes a useful module called SUPERVISION that is designed to split the input file into several subfiles and merge the split result files, thus reducing drastically the computation time by allowing manual start of parallel sessions. This module has however rarely been employed to date, possibly due to its laborious and time-demanding preparation procedure. This limitation is overcome in R.SamBada through the *sambadaParallel* function that implements SUPERVISION by default, and relies on the DOPARALLEL R-package (Microsoft Corporation & Weston, 2017). Furthermore, unlike the previous version of SAMβADA (0.5.1 used in Stucki et al., 2017), version 0.8.1 (included in R.SamBada) makes it possible to directly assess the effect of population structure by comparing the full model (containing all population variables and one or more environmental variables) with the null model (containing only population variables).

2.1.3 | Postprocessing

Four ad hoc functions have been developed for obtaining and visualizing SAMβADA's outputs. In the postprocessing pipeline, the statistical significance of genotype–environment associations is derived since only G- and Wald-scores are calculated by SAMβADA, and no hypothesis testing is performed. Here, R.SamBada provides the function *prepareOutput*, which computes (i) *p*-values by comparing the spread

of G- or Wald-scores from SAMβADA to a chi-squared distribution and (ii) *q*-values based on Storey's method (Storey, 2003). The visualization of the position of outlier loci along the genome is possible using the *plotManhattan* function that generates Manhattan plots based on the *p*- or *q*-values as computed by *prepareOutput*.

Next, *plotResultInteractive* can be used to display interactive Manhattan plots. In particular, users can specify which chromosome(s) they want to visualize for which environmental variable, the *p*- or *q*-values, being then plotted for each genotype as a function of their genomic coordinates. Marker name, position, *p*-value, functional relevance (e.g., intergenic-, nonsynonymous variants) as well as proximal genes – whenever present – can be then retrieved for each marker by directly clicking on the set of points of interest being displayed. Gene annotation and functional investigation are performed by internal calls to the Ensembl genome browser (Hubbard et al., 2002) and the Variant Effect Predictor (VEP) (Yates et al., 2015), respectively, while the whole interactive graphical interface relies on the R-package SHINY (Chang, Cheng, Allaire, Xie, & McPerson, 2018). Additionally, a basic geographic map shows the geographic distribution of the marker, the environmental variable and the population structure (examples presented in Figure S1).

Finally, the *plotMap* mapping function makes it possible to represent the geographic distribution of (a) the putative signature(s) of selection, (b) the environmental pressure associated (as a raster background if available), (c) the neutral population structure (Figure 5 for an example) and (d) the degree of genetic similarity among sampling sites for the target markers (i.e., its spatial autocorrelation, see Stucki et al., 2017). *plotMap* relies on the functionalities embedded within the PACKCIRCLES R-package (Bedward, Eppstein, & Menzel, 2018) to shift nearby sampling points and prevent them from overlapping.

2.2 | Case studies

2.2.1 | Moroccan sheep

Sampling and genetic data

Preprocessing

Quality control analysis was performed using the *prepareGeno* function with MAF <0.05 and SNP missingness <0.1, leading to a pruned data set composed by 20,226,452 SNPs (corresponding to 60,679,356 genotypes). SRTM and Worldclim variables (56 in

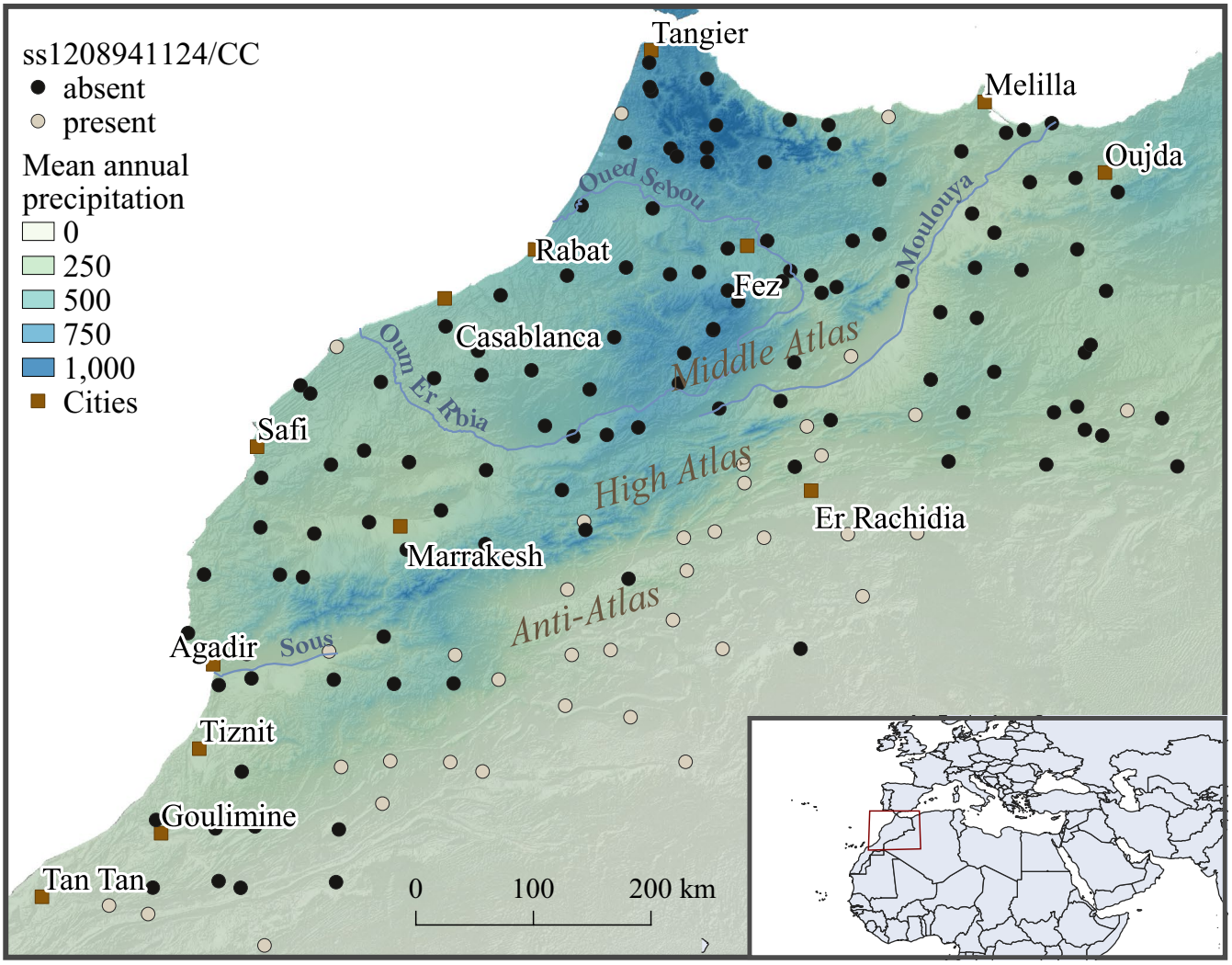

FIGURE 3 Spatial occurrence of the CC genotype for SNP ss1208941124. In the background, the shaded topography with mean annual precipitation (given in [mm/year]) is displayed [Colour figure can be viewed at wileyonlinelibrary.com]

total) were downloaded with *createEnv*, and *prepareEnv* was run to check for variable correlation in order to exclude variables showing an r^2 higher than 90%, resulting in a final data set consisting of 16 environmental variables (13 Bioclim variables, 2 raw WorldClim and altitude). No population variable was included in SAMβADA's models (univariate mode) since no evidence of population structure emerged using the PCA method implemented in SNPRELATE (with genomic filter of MAF <0.05, SNP missingness <0.1 and LD threshold <0.2).

Postprocessing

q-values based on G-scores were visualized with a Manhattan plot using a significance threshold of 0.05. *plotResultInteractive* was used to detect genes neighbouring the markers under selection as well as to identify variant functions (e.g., nonsynonymous SNPs).

2.3 | Spanish Lidia cattle

2.3.1 | Sampling and genetic data

The Lidia cattle breed (*Bos taurus*) emerged during the XVIII century and evolved mainly in the *dehesas* ecosystems of the west/south-west Iberian Peninsula, composed of pasturelands interspersed with Mediterranean oaks (*Quercus ilex*) (del Barrio, Ponce, Benavides, & Roig, 2014). Since its establishment, Lidia was prompt to isolation by preventing crossbreeding with allochthonous cattle (Eusebi, Cortés, Dunner, & Cañón, 2017) and became fragmented into reproductively isolated lineages (called *encastes*) with homogeneous morphology, behaviour and genetics (Boletin Oficial del Estado, 2001). Such a peculiar evolutionary and cultural context boosted Lidia's population size to become the largest Spanish breed and made it one of the most inclusive intergrading bovine

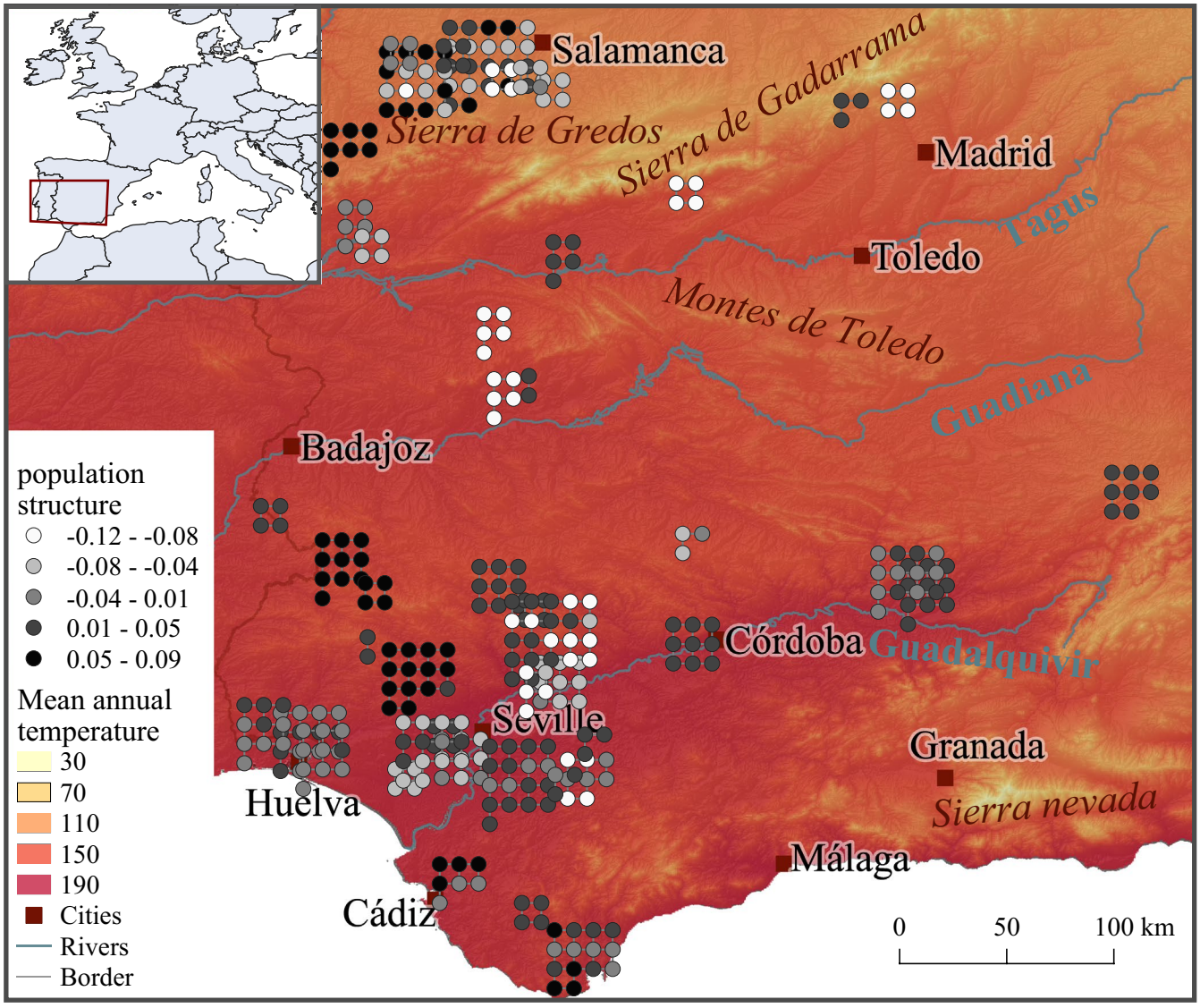

FIGURE 4 Spatial distribution of the Lidia cattle population structure according to the scores of the first principal component, with a shaded relief and mean annual temperature [°C * 10] as background, as provided in the WorldClim database. Due to overlaps, close points are scattered around the farm [Colour figure can be viewed at wileyonlinelibrary.com]

population, granting high level of genetic richness among *encastes* coupled with low average genetic diversity values within lineages (Cañón et al., 2008). A total of 349 individuals were sampled among 61 different breeders evenly distributed across southern Spain's *dehesas* region (Figure 4). Between one and seventeen animals per breeder were selected based on pedigree information to minimize the risk of kinship among individuals. Animals were genotyped using the Illumina BovineSNP50 array v.2 (Eusebi et al., 2017).

2.3.2 | Preprocessing

Quality control analysis was performed using the *prepareGeno* function with a MAF <0.05 and SNP missingness <0.1. The resulting molecular data set consisted of 38,335 SNPs (i.e., 115,005 genotypes). SRTM and Worldclim variables (56 in total) were downloaded with the *createEnv* function, and *prepareEnv* was used to test for variable correlation resulting in only 15 variables (10 Bioclim and 5 raw WorldClim variables) kept which showed a r^2 lower than 90%. Due to the presence of population structure observed with SNPRELATE's PCA method (see Results section), SAMβADA was run in bivariate mode by adding a variable to account for population structure (score of the first PCA). This variable is not correlated with other kept environmental factors (highest correlation: precipitation in April, $r^2 = 0.25$).

2.3.3 | Postprocessing

p-values based on *G*-Scores were corrected for multiple testing with Bonferroni method and subsequently were displayed in a Manhattan plot (*q*-values were not conservative enough in that case), with a significance threshold of 0.05, and *plotResultInteractive* was then used to detect associated genes.

3 | RESULTS

3.1 | Time efficiency

Besides the time saved during pre- and postprocessing, R.SamBada is more time-efficient than using SAMβADA's command line (version 0.5.1) for two reasons: first, R.SamBada automatically integrates SUPERVISION to distribute the processing of models over several cores, which makes the analysis run *x* times faster (where *x* represents the number of CPU), to which we must add a few minutes to split and merge the data set (e.g., 24 min to split and merge the sheep data set, compared to 160 hr saved by parallel computing on the same 11 cores). Second, if population variables are included in the analysis, the new version of SAMβADA (0.8.1) will only focus on models including population variables. Here, the time saved will depend on the number of population variables (for the Lidia cattle analysis, with one population variable, it reduced the computing time from 53 to 9 min).

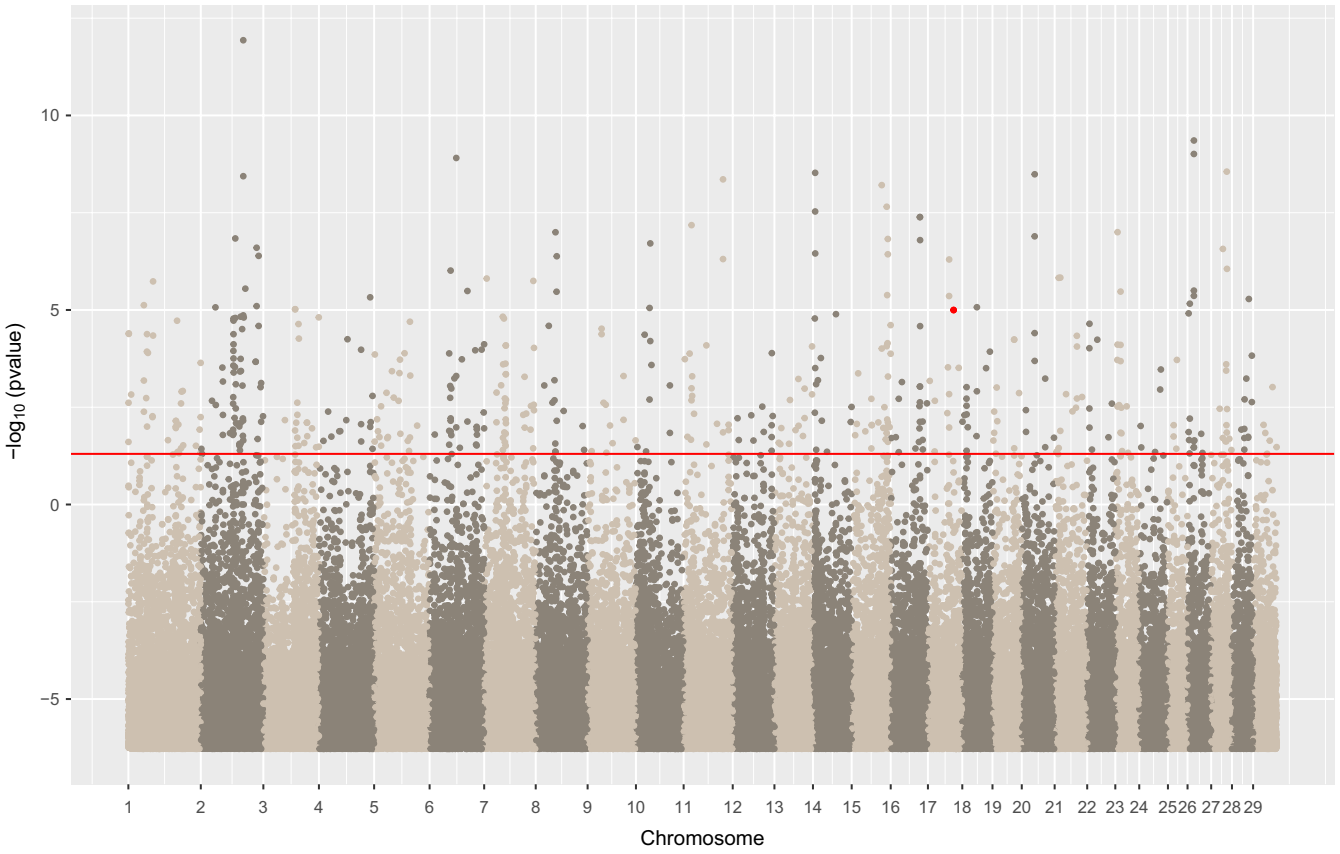

FIGURE 5 Manhattan plot of the Lidia cattle study, showing the *p*-values with Bonferroni correction as derived from the SAMβADA models involving mean annual temperature and one population variable. The red point corresponds to SNP ARS-BFGL-NGS-106879, located 30,000 base pairs apart from the HSPB8 gene [Colour figure can be viewed at wileyonlinelibrary.com]

3.2 | Moroccan sheep

3.2.1 | Population structure

The variance explained by the first three PCA components was 0.0085, 0.0083 and 0.0082, respectively, indicating no clear population structure. Therefore, no variable translating population structure was retained for subsequent analyses.

3.2.2 | Genotype-environment associations

When investigating SAMβADA's results, a significant peak around position 4.38e7 was observed on chromosome 23 in association with annual precipitation (Figure 2). Within this genomic region, two SNPs (i.e., ss1208941124 at position 23: 43867891 and ss1208941157 at position 23: 43869831) were found to be nonsynonymous for the gene *MC5R* (melanocortin 5 receptor) and in strong LD (r^2 = 0.97).

Given such a high LD, the spatial distribution of these markers is almost identical (except for one individual; data not shown), and only ss1208941124 is illustrated (Figure 3). For this locus, genotype CC is very frequent in the northern part of Morocco, where annual precipitation is on average high (reaching values of 1,000 mm/year), while being almost absent in the south (at the Sahara Desert's gate where precipitation is as low as 50 mm/year).

3.3 | Lidia cattle in Spain

3.3.1 | Population structure

The variance explained by the first three components of the PCA was 0.049, 0.029 and 0.024, respectively. In this case, the first principal component is likely to represent population structure, given the difference in variance observed between PC 1 and 2,

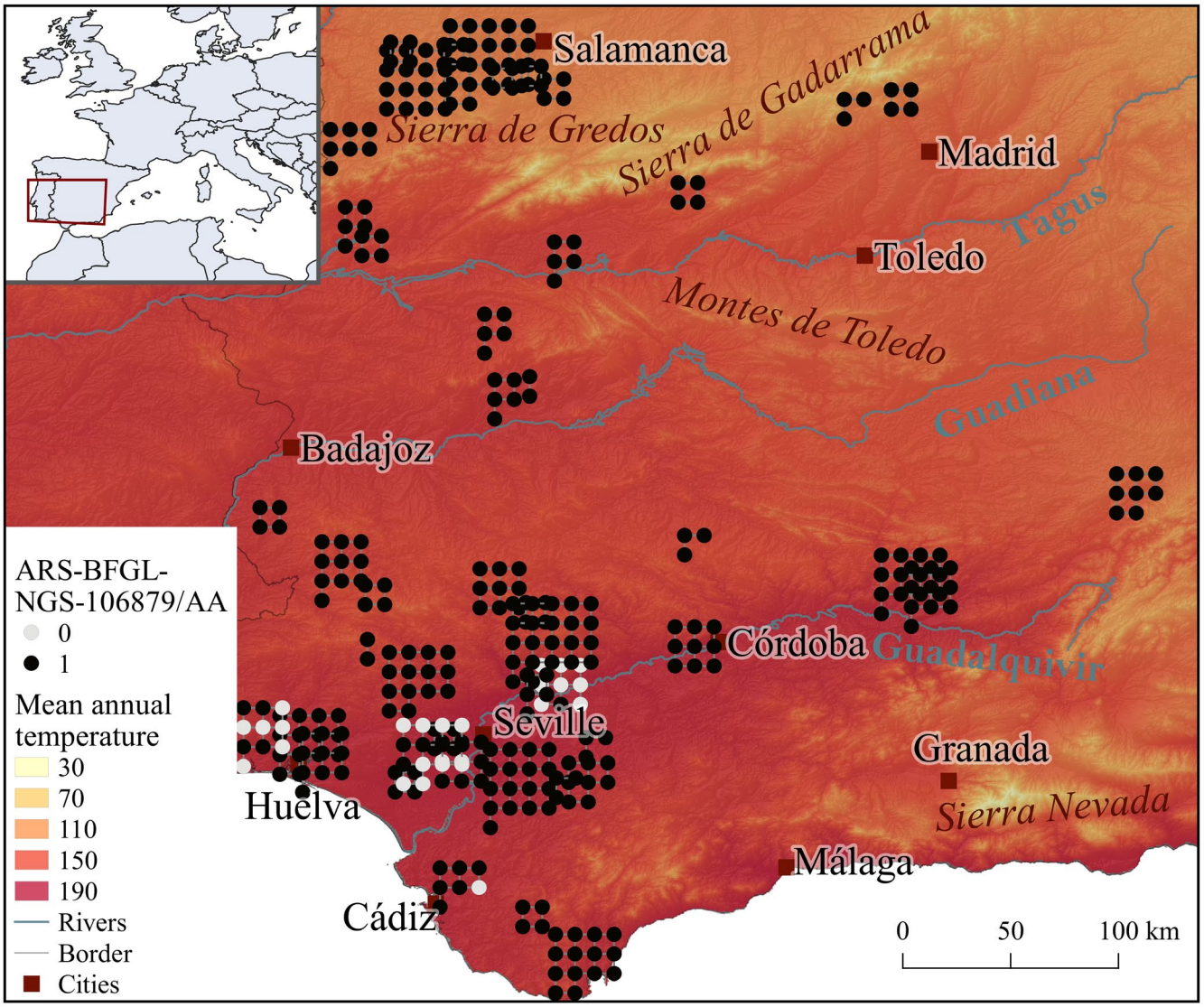

FIGURE 6 Presence–absence of the AA genotype of SNP ARS-BFGL-NGS-106879 reported with shaded relief and mean annual temperature [°C * 10]) as background. Due to overlaps, close points are scattered around the farm [Colour figure can be viewed at wileyonlinelibrary.com]

and in accordance with what has been previously observed in between European cattle breeds (e.g., Orozco-terWengel et al., 2015). Geographically, genetic clusters composed of either single or groups of proximately located farms were identified (e.g., south from Badajoz), although no wider spatial pattern was evident (e.g., north–south gradient) (Figure 4).

3.3.2 | Genotype-environment associations

Several narrow peaks were observed in the models involving mean annual temperature (i.e., bio1 bioclim variable) (Figure 5). In particular, the Ensembl query revealed the SNP ARS-BFGL-NGS-106879 (at position 17: 56127482) to be located ~ 30,000 base pairs from the gene *HSPB8* (heat shock protein family B [small] member 8).

Spatial occurrence of genotype AA from ARS-BFGL-NGS-106879 appears to be related to mean annual temperature (Figure 6). More specifically, this genotype is geographically widespread in the study area, except for 23 individuals found in different farms from the Guadalquivir valley, a region with temperature reaching 36°C during the hottest month of the year. Importantly, however, when comparing Figures 4 and 6 it can be seen that the genotype distribution does not match the prevailing population structure; hence, this result is independent of the calculated population structure present within the breed.

4 | DISCUSSION

4.1 | Role of the package

We have provided a demonstration of R.SamBada, encompassing the entire pipeline analysis from *pre-* to *post hoc* processing, following the classical SAMβADA analysis pathway, but much more efficiently. R.SamBada helps saving user's time for preparing input files thanks to newly built functions, as well as computing time through better integration of population structure and automated split of computations on parallel cores. Additionally, it provides a standardized processing chain, thus facilitating reproducibility.

Moreover, part of the pre- and postprocessing chain can possibly be coupled with other software used in landscape genomics and more generally with software designed to detect signature of selection. For example, the postprocessing function *plotResultInteractive* could be used with any type of outputs as long as its structure is similar to the returned value of *prepareOutput* (i.e., columns indicating the position of the SNP as well as the p-value associated with the corresponding genotype; refer to the package documentation for more detail).

4.2 | Case studies

4.2.1 | Sheep in Morocco

Two of the SNPs on chromosome 23 associated with precipitation (ss1208941124 and ss1208941157) are nonsynonymous variants located within the *MC5R* gene. Although understudied in sheep, this gene has been reported to be linked to a wide range of physiological functions in different mammal species, including regulation of food intake and sebum secretion (Switonski, Mankowska, & Salamon, 2013). Wax secretion is of particular interest with respect to precipitation; indeed, sebaceous secretions in Merino sheep have been found to hinder *Dermatophilus dermatonomous* infection (Roberts, 1963), a skin disease affecting many domestic and wild animal species that can be lethal in extreme cases. In the same breed, Dermatophilosis outbreaks have been found to be linked with exceptionally rainy years (Yeruham, Elad, & Nyska, 1995). Thus, the secretion of wax could play an important role in protecting sheep against rainy weather, consistent with its environmental relationship with annual precipitation here.

4.2.2 | Lidia cattle

The SNP ARS-BFGL-NGS-106879 is associated with mean annual temperature and located in the vicinity of the gene *HSPB8*. This gene is thought to code for a chaperone protein, which is upregulated in presence of heat and other environmental stress, and exerts an important cytoprotective role (Verma et al., 2016). In cattle, this gene was found to be associated with heat tolerance in both crossbred and pure *Bos indicus* Sahiwal in India (Sengar et al., 2018; Verma et al., 2016) that can suggest its putative involvement with adaptation to heat tolerance in Lidia cattle as well.

This SNP lies at ~30 Kbp outside the *HSPB8* coding region, either suggesting the SNP to be in LD with some adaptive variant within the gene or to possibly have an important regulatory effect on transcription. However, considering the relatively low average LD between loci at 30Kbp-distance (computed r^2 in this region = 0.2), the existence of a significant variant within the gene is unlikely. In contrast, such a distance would suggest more likely this SNP to be involved in regulatory processes; indeed, according to Brodie, Azaria, and Ofran (2016), large insertions/deletions with regulative roles can be found as far as 2Mbp around a gene and associated with nearby SNPs.

4.3 | Perspectives

R.SamBada represents a step forward in facilitating the chain of processes required to implement a landscape genomics study. However, several further improvements could be implemented in the future. For example, the query based on the Ensembl database requires a reference genome for the species under investigation, which remains relatively uncommon for nonmodel species. It would therefore be very useful to further develop functions performing a BLAST alignment (Johnson et al., 2008) and see if any match can be found with orthologous genes from related species where genomes have been produced.

In addition, functionalities could be augmented to help the user define ad hoc QC thresholds. For instance, a function allowing species-specific estimation of LD in order to better calibrate the pruning applied before computing the PCA would be useful. Furthermore, R.SamBada currently only implements basic QC of genetic data

(MAF, LD, missingness) and does not test for other useful checks (e.g., Identity By Descent – IBD – or Hardy–Weinberg Equilibrium – HWE). However, such controls can easily be performed with dedicated software like PLINK (C.C. Chang et al., 2015) or vcftools (Danecek et al., 2011) before entering SAMβADA's R-pipeline. Moreover, SAMβADA is one among several software solutions to detect selection signatures in a spatial context and can be used in combination with other packages like LFMM (Caye et al., 2019), BayEnv (Günther & Coop, 2013) or both (Stucki et al., 2017) in order to compare the results obtained. Further functionalities could be developed to ease the computation and comparison with those methods.

Finally, it is important to keep in mind that landscape genomic approaches such as SAMβADA implement an explanatory analysis which allows rapid identification of candidate genes, but lacks a validation procedure, meaning that derived hypotheses need to be further tested (e.g., through investigation of variant effect on protein tertiary structure and function or through laboratory experiments).

ACKNOWLEDGEMENTS

This work was supported by the European Union 7th framework project NEXTGEN (Grant Agreement no. 244356, coordinated by P.T.) and the FACCE ERA-NET Plus project CLIMGEN (grant ANR-14-JFAC-0002-01). MWB and POTW were funded by BBSRC through the FACCE-JPI ERA-NET Climate Smart Agriculture project CLIMGEN (BB/M019276/1). NS is a recipient of a Marie Skłodowska-Curie Individual Fellowship funded by the European Union's Horizon 2020 research and innovation programme under grant agreement No DLV-655100.

AUTHOR CONTRIBUTIONS

S.Dur. wrote the major part of the R-package with the help of O.S., E.V. and S.S. on specific points. In particular, the new functionalities of the C++ code were developed by S.S. N.S. wrote the sections of the manuscript dedicated to the Lidia cattle case study with the help of S.Dun. who was responsible of the Lidia samples availability. K.L. performed most of the analysis related to the Moroccan sheep case study and elaborated part of the related text. S.Dur. wrote the rest of the manuscript with the help of all authors. S.J. conceived and supervised the project. P.O., M.W.B. and S.J. revised the manuscript.

SOFTWARE AVAILABILITY

R.SamBada package is available in the R CRAN package repository and on GitHub (github.com/SolangeD/R.SamBada).

DATA ACCESSIBILITY

The Moroccan sheep data set is available at https://projects.ensembl.org/nextgen/ population MODA. The Lidia cattle data set is accessible from FigShare: https://doi.org/10.6084/m9.figshare.5394895.v4 (only Spanish samples included in the analysis).

ORCID

Solange Duruz https://orcid.org/0000-0002-1235-6747
Natalia Sevane https://orcid.org/0000-0003-4766-6291
Oliver Selmoni https://orcid.org/0000-0003-0904-5486
Elia Vajana https://orcid.org/0000-0003-1340-3389
Kevin Leempoel https://orcid.org/0000-0001-7335-7930
Sylvie Stucki https://orcid.org/0000-0002-1624-1576
Pablo Orozco-terWengel https://orcid.org/0000-0002-7951-4148
Estelle Rochat https://orcid.org/0000-0002-7978-5239
Susana Dunner https://orcid.org/0000-0001-8637-2208
Michael W. Bruford https://orcid.org/0000-0001-6357-6080
Stéphane Joost https://orcid.org/0000-0002-1184-7501

REFERENCES

Alberto, F. J., Boyer, F., Orozco-terWengel, P., Streeter, I., Servin, B., de Villemereuil, P., … Pompanon, F. (2018). Convergent genomic signatures of domestication in sheep and goats. *Nature Communications*, *9*(1), 813. https://doi.org/10.1038/s41467-018-03206-y

Alexander, D. H., Novembre, J., & Lange, K. (2009). Fast model-based estimation of ancestry in unrelated individuals. *Genome Research*, *19*(9), 1655–1664. https://doi.org/10.1101/gr.094052.109

Balkenhol, N., Dudaniec, R. Y., Krutovsky, K. V., Johnson, J. S., Cairns, D. M., Segelbacher, G. … Joost, S. (2019). Landscape genomics: Understanding relationships between environmental heterogeneity and genomic characteristics of populations. In O. P. Rajora (Ed.), *Population genomics: Concepts, approaches and applications*, *population genomics* (pp. 261–322). Cham, Switzerland: Springer International Publishing. https://doi.org/10.1007/13836_2017_2

Bedward, M., Eppstein, D., & Menzel, P. (2018). *packcircles: Circle packing*. Retrieved from https://CRAN.R-project.org/package=packcircles

Boletin Oficial del Estado (2001). *Boletin Oficial del Estado*. REAL DECRETO 60/2001, de 26 de enero, sobre prototipo racial de la raza bovina de lidia. pp. 5255–5261.

Brodie, A., Azaria, J. R., & Ofran, Y. (2016). How far from the SNP may the causative genes be? *Nucleic Acids Research*, *44*(13), 6046–6054. https://doi.org/10.1093/nar/gkw500

Cañón, J., Tupac-Yupanqui, I., García-Atance, M. A., Cortés, O., Garcia, D., Fernández, J., & Dunner, S. (2008). Genetic variation within the Lidia bovine breed. *Animal Genetics*, *39*(4), 439–445. https://doi.org/10.1111/j.1365-2052.2008.01738.x

Caye, K., Jumentier, B., Lepeule, J., & François, O. (2019). LFMM 2: Fast and accurate inference of gene-environment associations in genome-wide studies. *Molecular Biology and Evolution*. https://doi.org/10.1093/molbev/msz008

Cesconeto, R. J., Joost, S., McManus, C. M., Paiva, S. R., Cobuci, J. A., & Braccini, J. (2017). Landscape genomic approach to detect selection signatures in locally adapted Brazilian swine genetic groups. *Ecology and Evolution*, *7*(22), 9544–9556. https://doi.org/10.1002/ece3.3323

Chang, C. C., Chow, C. C., Tellier, L. C., Vattikuti, S., Purcell, S. M., & Lee, J. J. (2015). Second-generation PLINK: Rising to the challenge of larger and richer datasets. *Gigascience*, *4*(1), 1. https://doi.org/10.1186/s13742-015-0047-8

Chang, W., Cheng, J., Allaire, J., Xie, Y., & McPerson, J. (2018). *shiny: Web Application Framework from R*. Retrieved from https://CRAN:R-project.org/package=shiny

Corporation, M., & Weston, S. (2017). *doParallel: Foreach parallel adaptor for the "parallel" package*. Retrieved from https://CRAN.R-project.org/package=doParallel

Cuervo-Alarcon, L., Arend, M., Müller, M., Sperisen, C., Finkeldey, R., & Krutovsky, K. V. (2018). Genetic variation and signatures of natural selection in populations of European beech (*Fagus sylvatica* L.) along precipitation gradients. *Tree Genetics & Genomes*, *14*(6), 84. https://doi.org/10.1007/s11295-018-1297-2

Danecek, P., Auton, A., Abecasis, G., Albers, C. A., Banks, E., DePristo, M. A., … Durbin, R. (2011). The variant call format and VCFtools. *Bioinformatics*, *27*(15), 2156–2158. https://doi.org/10.1093/bioinformatics/btr330

del Barrio, J. M. G., Ponce, R. A., Benavides, R., & Roig, S. (2014). Species richness of vascular plants along the climatic range of the Spanish dehesas at two spatial scales. *Forest Systems*, *23*(1), 111–119. https://doi.org/10.5424/fs/2014231-04521

Eusebi, P. G., Cortés, O., Dunner, S., & Cañón, J. (2017). Genomic diversity and population structure of Mexican and Spanish bovine Lidia breed. *Animal Genetics*, *48*(6), 682–685. https://doi.org/10.1111/age.12618

Farr, T. G., Rosen, P. A., Caro, E., Crippen, R., Duren, R., Hensley, S., … Alsdorf, D. (2007). The shuttle radar topography mission. *Reviews of Geophysics*, *45*, RG2004. https://doi.org/10.1029/2005RG000183

Guessous, F., Boujenane, I., Bourfia, M., & Narjisse, H. (1989). *Sheep in Morocco*. FAO Animal Production and Health Paper (FAO).

Günther, T., & Coop, G. (2013). Robust identification of local adaptation from allele frequencies. *Genetics*, *195*(1), 205–220. https://doi.org/10.1534/genetics.113.152462

Hijmans, R. J., Cameron, S. E., Parra, J. L., Jones, P. G., & Jarvis, A. (2004). *The WorldClim interpolated global terrestrial climate surfaces*. Version 1.3.

Hubbard, T., Barker, D., Birney, E., Cameron, G., Chen, Y., Clark, L., 2002). The Ensembl genome database project. *Nucleic Acids Research*, *30*(1), 38–41. https://doi.org/10.1093/nar/30.1.38

Johnson, M., Zaretskaya, I., Raytselis, Y., Merezhuk, Y., McGinnis, S., & Madden, T. L. (2008). NCBI BLAST: A better web interface. *Nucleic Acids Research*, *36*(suppl_2), W5–W9. https://doi.org/10.1093/nar/gkn201

Joost, S., Bonin, A., Bruford, M. W., Després, L., Conord, C., Erhardt, G., & Taberlet, P. (2007). A spatial analysis method (SAM) to detect candidate loci for selection: Towards a landscape genomics approach to adaptation. *Molecular Ecology*, *16*, 3955–3969.

Kawecki, T. J., & Ebert, D. (2004). Conceptual issues in local adaptation. *Ecology Letters*, *7*(12), 1225–1241. https://doi.org/10.1111/j.1461-0248.2004.00684.x

Lee, C., Abdool, A., & Huang, C.-H. (2009). PCA-based population structure inference with generic clustering algorithms. *BMC Bioinformatics*, *10*(1), S73. https://doi.org/10.1186/1471-2105-10-S1-S73

Leempoel, K., Duruz, S., Rochat, E., Widmer, I., Orozco-terWengel, P., & Joost, S. (2017). Simple rules for an efficient use of geographic information systems in molecular ecology. *Frontiers in Ecology and Evolution*, *5*, 33. https://doi.org/10.3389/fevo.2017.00033

Lucek, K., Keller, I., Nolte, A. W., & Seehausen, O. (2018). Distinct colonization waves underlie the diversification of the freshwater sculpin (*Cottus gobio*) in the Central European Alpine region. *Journal of Evolutionary Biology*, *31*, 1254–1267.

Orozco-terWengel, P., Barbato, M., Nicolazzi, E., Biscarini, F., Milanesi, M., Davies, W., … Bruford, M. W. (2015). Revisiting demographic processes in cattle with genome-wide population genetic analysis. *Frontiers in Genetics*, *6*, 191. https://doi.org/10.3389/fgene.2015.00191

Pereira, P., Teixeira, J., & Velo-Antón, G. (2018). Allele surfing shaped the genetic structure of the European pond turtle via colonization and population expansion across the Iberian Peninsula from Africa. *Journal of Biogeography*, *45*(9), 2202–2215. https://doi.org/10.1111/jbi.13412

R Core Team (2018). *R: A Language and Environment for Statistical Computing*. Retrieved from https://www.R-project.org/

Rellstab, C., Gugerli, F., Eckert, A. J., Hancock, A. M., & Holderegger, R. (2015). A practical guide to environmental association analysis in landscape genomics. *Molecular Ecology*, *24*(17), 4348–4370. https://doi.org/10.1111/mec.13322

Roberts, D. S. (1963). Barriers to Dermatopilus dermatonomus infection on the skin of sheep. *Australian Journal of Agricultural Research*, *14*(4), 492–508. https://doi.org/10.1071/AR9630492

Sengar, G. S., Deb, R., Singh, U., Raja, T. V., Kant, R., Sajjanar, B., … Joshi, C. G. (2018). Differential expression of microRNAs associated with thermal stress in Frieswal (Bos taurus x Bos indicus) crossbred dairy cattle. *Cell Stress and Chaperones*, *23*(1), 155–170. https://doi.org/10.1007/s12192-017-0833-6

Shih, K.-M., Chang, C.-T., Chung, J.-D., Chiang, Y.-C., & Hwang, S.-Y. (2018). Adaptive genetic divergence despite significant isolation-by-distance in populations of taiwan cow-tail fir (*Keteleeria davidiana* var. *formosana*). *Frontiers in Plant Science*, *9*, 92. https://doi.org/10.3389/fpls.2018.00092

Storey, J. D. (2003). The positive false discovery rate: A Bayesian interpretation and the q-value. *The Annals of Statistics*, *31*(6), 2013–2035. https://doi.org/10.1214/aos/1074290335

Stucki, S., Orozco-terWengel, P., Forester, B. R., Duruz, S., Colli, L., Masembe, C.…Consortium, N. (2017). High performance computation of landscape genomic models including local indicators of spatial association. *Molecular Ecology Resources*, *17*(5), 1072–1089. https://doi.org/10.1111/1755-0998.12629

Switonski, M., Mankowska, M., & Salamon, S. (2013). Family of melanocortin receptor (MCR) genes in mammals—mutations, polymorphisms and phenotypic effects. *Journal of Applied Genetics*, *54*(4), 461–472. https://doi.org/10.1007/s13353-013-0163-z

Vajana, E., Barbato, M., Colli, L., Milanesi, M., Rochat, E., Fabrizi, E., … Ajmone-Marsan, P. (2018). Combining landscape genomics and ecological modelling to investigate local adaptation of Indigenous Ugandan Cattle to East Coast Fever. *Frontiers in Genetics*, *9*, 385. https://doi.org/10.3389/fgene.2018.00385

Verma, N., Gupta, I. D., Verma, A., Kumar, R., Das, R., & M.R., V., (2016). Novel SNPs in HSPB8 gene and their association with heat tolerance traits in Sahiwal indigenous cattle. *Tropical Animal Health and Production*, *48*(1), 175–180. https://doi.org/10.1007/s11250-015-0938-9

Yates, A., Beal, K., Keenan, S., McLaren, W., Pignatelli, M., Ritchie, G. R. S., … Flicek, P. (2015). The Ensembl REST API: Ensembl data for any language. *Bioinformatics*, *31*(1), 143–145. https://doi.org/10.1093/bioinformatics/btu613

Yeruham, I., Elad, D., & Nyska, A. (1995). Skin diseases in a Merino sheep herd related to an excessively rainy winter in a Mediterranean climatic zone. *Journal of Veterinary Medicine Series A*, *42*(1–10), 35–40. https://doi.org/10.1111/j.1439-0442.1995.tb00353.x

Zheng, X., Levine, D., Shen, J., Gogarten, S. M., Laurie, C., & Weir, B. S. (2012). A high-performance computing toolset for relatedness and principal component analysis of SNP data. *Bioinformatics*, *28*(24), 3326–3328. https://doi.org/10.1093/bioinformatics/bts606

SUPPORTING INFORMATION

Additional supporting information may be found online in the Supporting Information section at the end of the article.

How to cite this article: Duruz S, Sevane N, Selmoni O, et al; The NEXTGEN Consortium, The CLIMGEN Consortium.Rapid identification and interpretation of gene–environment associations using the new R.SamBada landscape genomics pipeline. *Mol Ecol Resour.* 2019;19:1355–1365. https://doi.org/10.1111/1755-0998.13044